Pictograph-to-Text Translation for Augmented and Alternative Communication

Leen Sevens*, Vincent Vandeghinste*, Lyan Verwimp, Ineke Schuurman*, Patrick Wambacq**, Frank Van Eynde***
Centre for Computational Linguistics (KU Leuven)
Department of Electrical Engineering (ESAT) (KU Leuven)
firstname.lastname@kuleuven.be

In today's digital age, people with limited reading and writing skills have trouble partaking in online activities. Not being able to access or use information technology is a major form of social exclusion. We present a Pictograph-to-Text translation system for people with an intellectual disability. It provides help in constructing Dutch textual messages, by allowing the user to input a series of pictographs, and translates these messages into natural language text.

The main challenge in translating from pictograph languages to natural language text is the fact that a pictograph-for-word correspondence will almost never provide an acceptable output. Pictographs are underspecified, both semantically and grammatically. In the second place, the pictograph input to translation could be ambiguous and unpredictable with respect to pictograph order.

Our baseline system for Pictograph-to-Text translation (Sevens et al. 2015) generates natural language from pictographs using language models and does not use any grammatical information in the translation process. When a pictograph is selected, its connected WordNet synset is retrieved, and from this synset, the system retrieves all the synonyms it contains. For each of these synonyms, reverse lemmatisation is applied. The reverse lemmatiser retrieves the full inflectional paradigm of each lemma. Each of these surface forms is a hypothesis for the language model. We propose two types of language models. In our *n*-gram-based approach, the system performs beam search decoding on an *n*-gram language model ($n \leq 5$), trained with the CMU toolkit (Clarkson & Rosenfeld 1997) on a Dutch corpus of over 1100M tokens. In our Long Short-Term Memory-based approach, we train a language model with Tensorflow (Abadi et al. 2016) on the Flemish part of the CGN corpus (3.8M tokens) (Oostdijk et al. 2002) and re-rank the natural language hypotheses. The evaluations of the baseline system show that using language models for finding the most likely combination of textual representations is already an improvement over the initial baseline (i.e., pictograph file names), but there is ample room for improvement.

In recent experiments, we apply machine translation techniques. Since a parallel corpus of pictograph sequences and well-formed written Dutch text is not available, we explore different approaches toward the creation of a suitable parallel corpus. In our first approach, we automatically translate a large corpus of monolingual Dutch SoNaR subtitles (27.6M tokens) (Oostdijk et al. 2013) into pictographs using the Text-to-Pictograph translation tool (Vandeghinste et al. 2015). In our second approach, we lemmatise the subtitle corpus, and remove all words that are not content words, thus creating a source language corpus that resembles pictograph input. Our phrase-based statistical machine translation approach toward Pictograph-to-Text translation uses the Moses decoder (Koehn et al. 2007), while our neural machine translation approach makes use of the open-source system OpenNMT (Klein et al. 2017). We build different models using a variety of training conditions, including factored models that include part-of-speech and lemma information, and evaluate all systems using automated metrics and human evaluations (adequacy, fluency, and ranking). Our first experiments indicate that the machine translation approaches outperform the baseline system.

References

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu & Xiaoqiang Zheng. 2016. Tensorflow: Large-scale Machine Learning on Heterogeneous Systems. URL: https://www.tensorflow.org/.

Clarkson, Philip & Roni Rosenfeld. 1997. Statistical Language Modeling using the CMU-Cambridge Toolkit. In: *The proceedings of Eurospeech*, Rodhes, Greece, 2707–2710.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart & Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. Proceedings of ACL 2017.

Koehn, Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin & Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In The Annual Meeting of the Association for Computational Linguistics, demonstration session (ACL), Prague, Czech Republic.

Oostdijk, Nelleke, Wim Goedertier, Frank Van Eynde, Louis Boves, Jean-Pierre Martens, Michael Moortgat & Harald Baayen. 2002. Experiences from the Spoken Dutch Corpus Project. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC), 340–347. Las Palmas, Canary Islands: European Language Resources Association.

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste & Henk van den Heuvel. 2013. SoNaR User Documentation.

Sevens, Leen, Vincent Vandeghinste, Ineke Schuurman & Frank Van Eynde. 2015b. Natural Language Generation from Pictographs. In: *Proceedings of 15th European Workshop on Natural Language Generation (ENLG)*, Brighton, UK.

Vandeghinste, Vincent, Ineke Schuurman, Leen Sevens & Frank Van Eynde. 2015. Translating Text into Pictographs. Natural Language Engineering, 1–28.